Analisi di un Servizio di Bike Sharing

Marco Zanella

Università di Padova

marco.zanella.9@studenti.unipd.it

3 luglio 2015

Table of Contents

Motivazione

I servizi di bike-sharing generano enormi quantità di dati: ottimi per tecniche di *data-mining*.

Predire la quantità di utenti: migliore organizzazione delle rastrelliere e degli interventi di manutenzione.

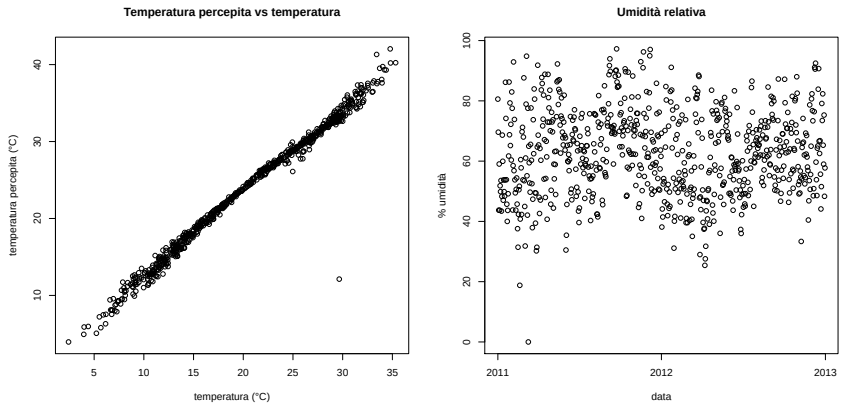
Hadi Fanaee-T and Joao Gama
Event labeling combining ensemble detectors and background knowledge

Analisi preliminare - I predittori

I predittori sono fattori esterni e/o ambientali:

- temperatura
- orario
- giorno festivo/feriale
- ...

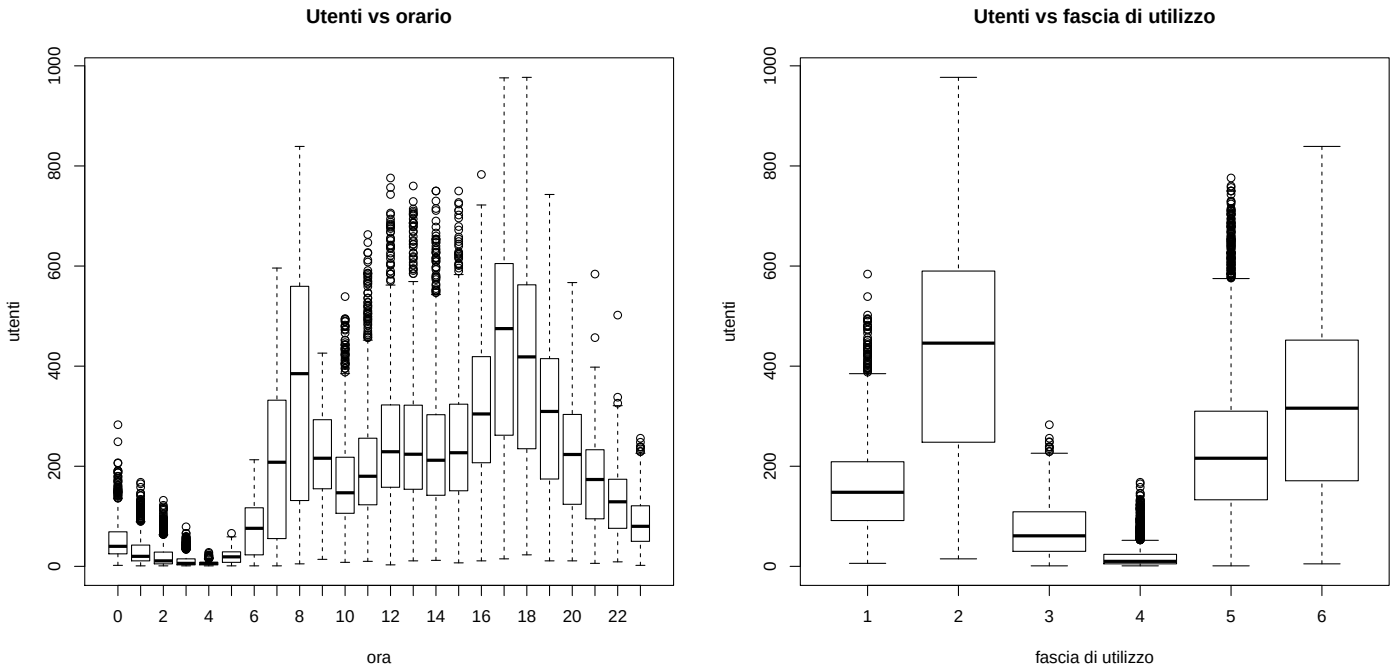
Possono avere importanza diversa, essere ridondanti o erronei:



Nel dataset, non ci sono informazioni mancanti.

Analisi preliminare - I predittori

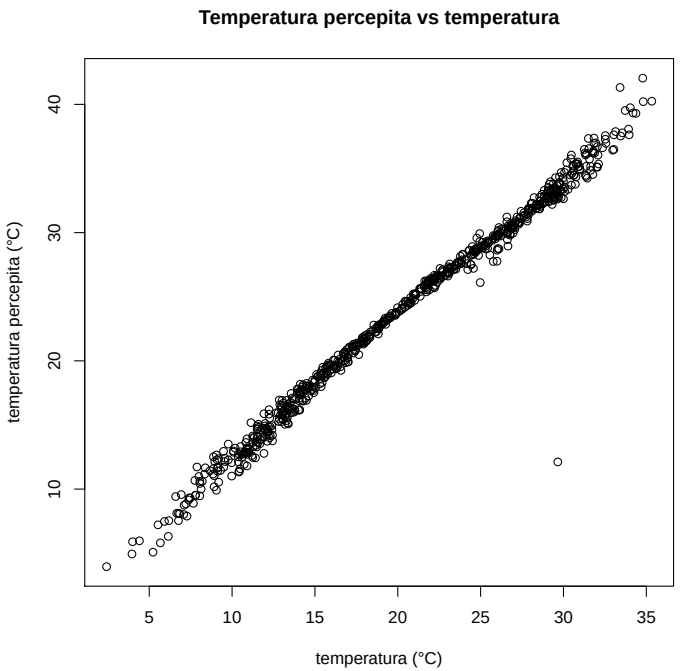
È possibile raggruppare alcune informazioni, come le ore:

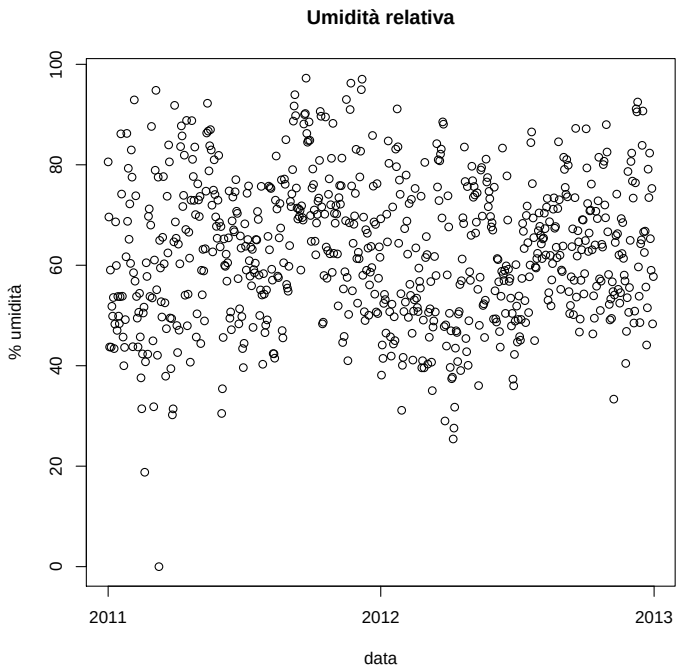


Gli algoritmi di *clustering* automatizzano il processo.

Analisi preliminare - Outlier

Punti sospetti, apparentemente anomali.





Richiedono un'investigazione manuale per valutarne mantenimento o rimozione.

Regressione - I modelli

I modelli sono costruiti su un *insieme di stima* (75% delle osservazioni):

- modello lineare
- MARS (con 1 e 2 gradi di interazione)
- GAM (con loess e splines)
- Projection Pursuit Regression
- Rete neurale
- Albero CART

Eventuali parametri sono stimati attraverso una scansione, suddividendo l'insieme di stima in *insieme di costruzione* e *di controllo*.

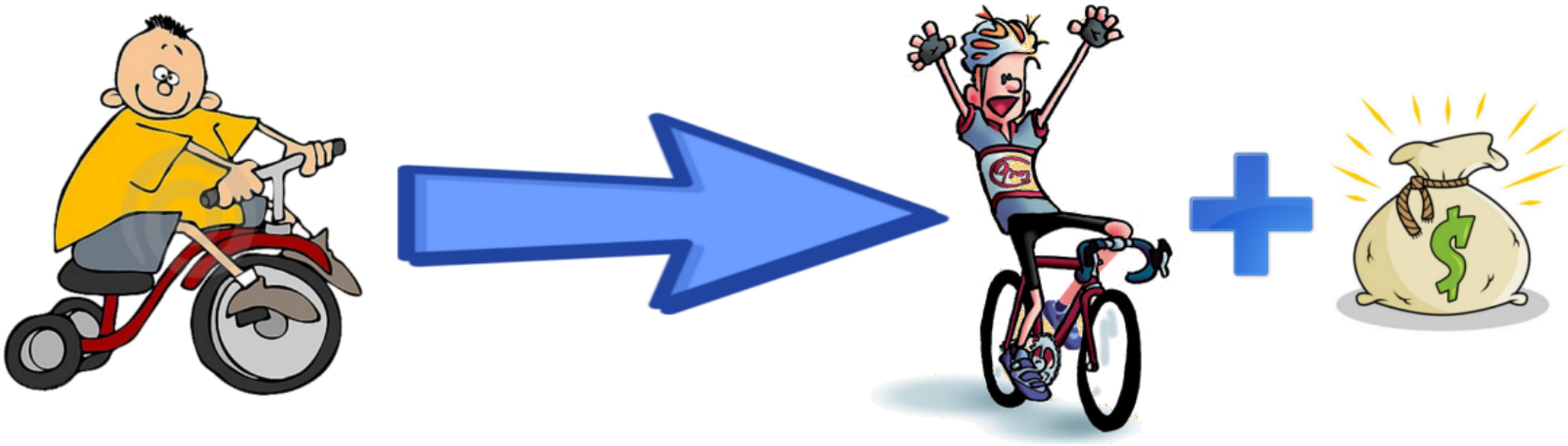
Regressione - Risultati

I risultati sono valutati in termini di MSE (*Mean Squared Error*, errore quadratico medio) sull'*insieme di verifica*.

modello	variabili	MSE
Lineare	tutte	12664.50
Lineare	solo sign.	13056.61
MARS	tutte (grado 1)	12619.53
MARS	tutte (grado 2)	6596.50
GAM (splines)	tutte	12643.32
GAM (splines)	solo sign.	12998.99
GAM (loess)	tutte	12602.14
GAM (loess)	solo sign.	12975.89
PPR	6 termini	5206.72
Rete neurale	-	9172.46
CART	50 foglie	7347.62
CART	13 foglie	10449.94

Classificazione

Quanti utenti sono occasionali? In quali circostanze sono più attivi?



Utente occasionale

Utente registrato + incassi

Gli utenti registrati portano guadagni maggiori: identificare gli utenti occasionali e convertirli in registrati.

Classificazione - I modelli

Modelli allenati sull'*insieme di stima*:

- modello lineare
- CART (classificazione)
- CART (regressione)
- MARS
- bagging
- boosting
- random forest

I modelli che producono valori *quantitativi* usano delle *soglie* per ottenere le *classi*.

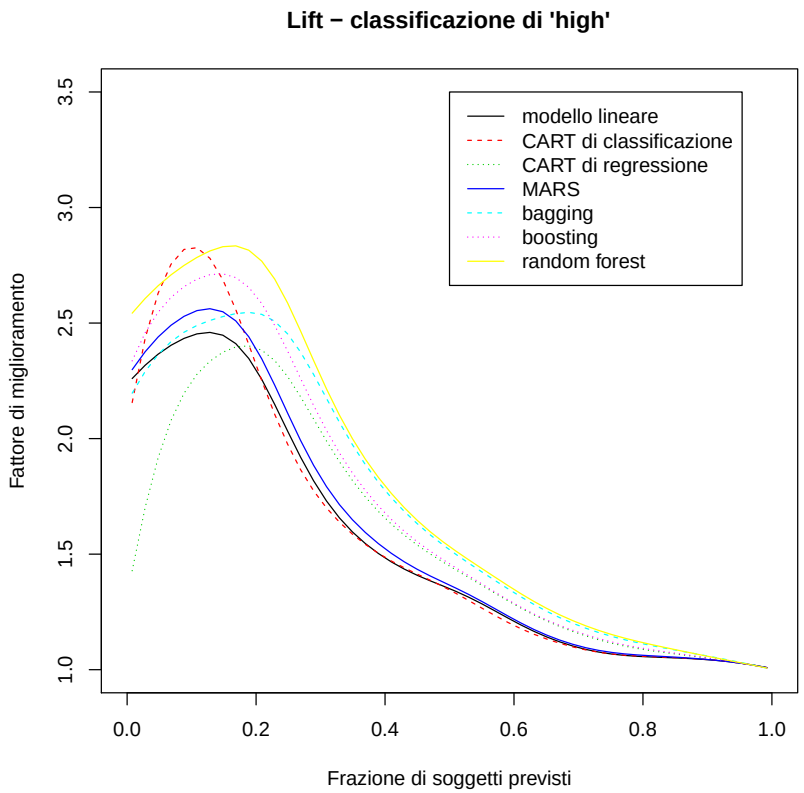
Classificazione - Risultati

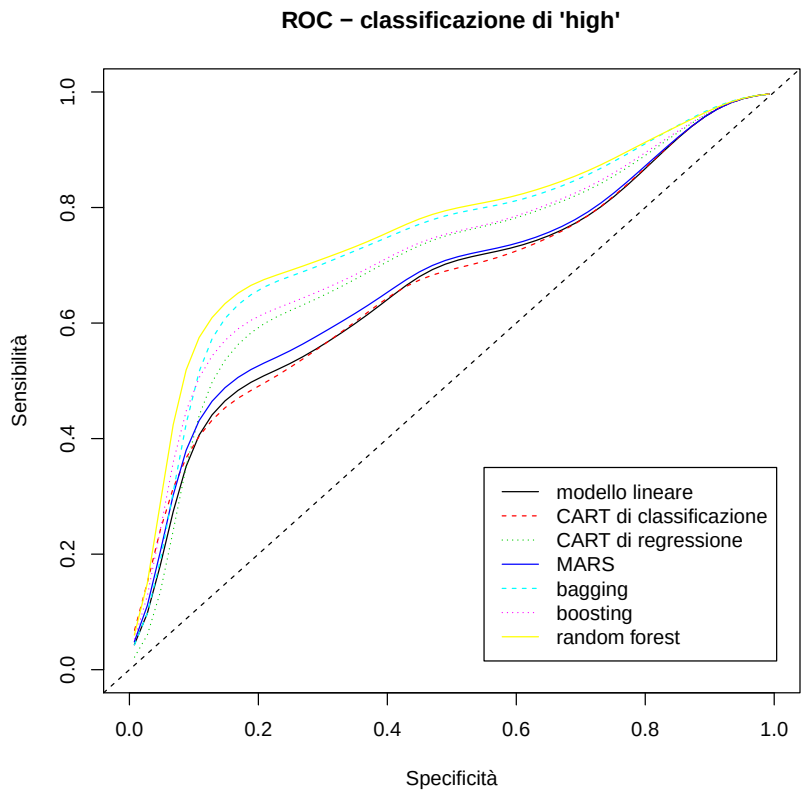
L'errore totale di classificazione viene usato come metrica.

Modello	errore
lineare	0.34
CART - classificazione	0.31
CART - regressione	0.31
MARS	0.34
bagging	0.29
boosting	0.28
random forest	0.26

Classificazione - Risultati

Gli errori sono molto simili: curve lift e ROC possono aiutare.

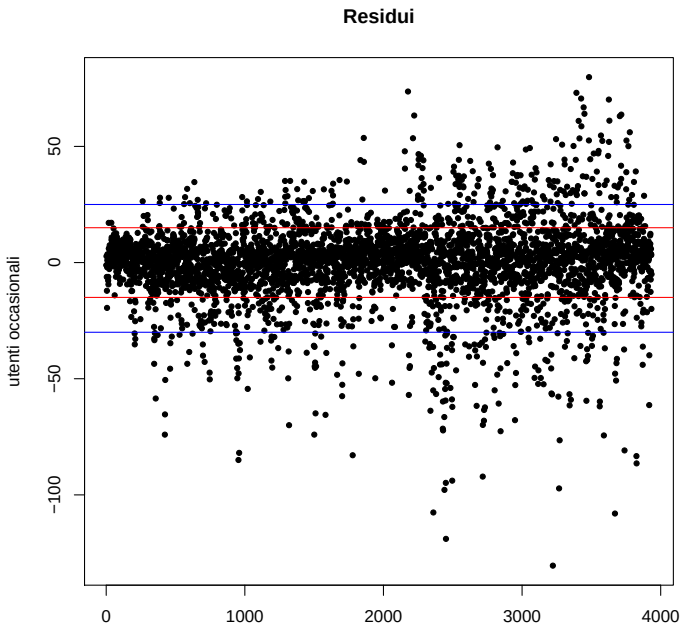




Il nostro interesse è maggiormente verso la classe *high*.

Classificazione - Risultati

Una piccola verifica informale: predire il numero di utenti casuali con i risultati ottenuti.



Residui *ben distribuiti*, $R^2 = 0.89$.

Conclusione – Grazie!

Conclusione - Regressione

- Analisi preliminare approfondita
- Diversi modelli esaminati e confrontati
- Scansioni esaustive per la stima dei parametri (regressione)
- Buoni risultati in termini di *MSE* (regressione)
- Buoni risultati in termini di errore, ROC e lift (classificazione)
- Verifica informale del classificatore positiva

Domande?